

Supplemental Analysis and Methods

Supplemental Analysis

iORF conservation and diversity among 26 wild yeast strains

We characterized iORF diversity in wild *S. paradoxus* populations. We used genomes from 24 strains that are structured in three main lineages named *SpA*, *SpB* and *SpC* (Charron et al. 2014; Leducq et al. 2016). Two *S. cerevisiae* strains were included as outgroups: the wild isolate YPS128 (Sniegowski et al. 2002; Peter et al. 2018) and the reference strain S288C. These lineages cover different levels of nucleotide divergence, ranging from ~ 13 % between *S. cerevisiae* and *S. paradoxus* to ~2.27 % between the two closest *SpB* and *SpC* lineages (Kellis et al. 2003; Leducq et al. 2016). We used microsynteny to identify and align orthologous non-genic regions between pairs of conserved annotated genes (Supplemental Fig. S1 and Methods). We identified 3,781 orthologous sets of intergenic sequences representing a total of ~ 2 Mb, with a median size of 381 bp. iORFs were annotated on aligned sequences using a method similar to the one employed by Carvunis et al. (2012), that is the first start and stop codons in the same reading frame not overlapping with known features, regardless of the strand, and with no minimum size. We then classified iORFs according to their conservation level among strains (Supplemental Fig. S1 and Methods). Because the annotation was performed on aligned sequences, we could precisely detect the presence/absence of orthologous iORFs among strains, based on the conservation of an iORF with the same start and stop positions without disruptive mutations in between. We used *S. cerevisiae* as an outgroup and removed iORFs present only in this species in order to focus on *S. paradoxus* diversity. However, we conserved iORFs present both in *S. cerevisiae* and in at least one *S. paradoxus* strain to consider the inter-species conservation.

We annotated 34,216 to 34,503 iORFs per *S. paradoxus* strain, for a total of 64,225 orthogroups annotated at least in one *S. paradoxus* strain (Supplemental Table S1). This represents a density of about 17 iORFs per Kb. The iORF set shows about 6 % conserved among *S. cerevisiae* and *S. paradoxus* strains, and 15 % specific and fixed within *S. paradoxus*. The remaining 79 % are still segregating within *S. paradoxus* (Supplemental Fig. S2A and Supplemental Table S1).

To understand how iORF diversity changes over a short evolutionary time scale, we estimated the age of iORFs and their turnover using ancestral sequence reconstruction (Fig. 1 and Methods). Because polymorphism within lineages (*SpA*, *SpB* or *SpC*) (Leducq et al. 2016) may affect the topology of the phylogeny (although most diversity in this group is among lineages), we used only one strain per lineage (YPS128 (*S. cerevisiae*), YPS744 (*SpA*), MSH-604 (*SpB*) and MSH-587-1 (*SpC*)) to reconstruct ancestral sequences at two divergence nodes that we labeled N1 for *SpB*-*SpC* divergence and N2 for *SpA*-*SpB/C* divergence. These strains contain 58,952 iORF orthogroups after removing the polymorphic iORFs that are absent in all the four selected strains. Reconstructed sequences were included in intergenic alignments of actual strains and were used to detect the presence or absence ancestral iORFs at each node (Supplemental Fig. S1 and Methods).

We estimated the age of the 58,952 iORFs and annotated the 2,291 iORFs detected only in ancestral sequences. 55 % of iORFs were present at N2 (the oldest age category) and are represented in each conservation group depending on iORF loss events occurring after N2 (Supplemental Fig. S2B and Table 1). We observed a continuous emergence of iORFs with 6,782 gains between N2 and N1, and 5,324 to 8,454 along terminal branches. We estimated a rate of emergence and loss at respectively 0.28 +/- 0.01 and 0.27 +/- 0.008 ORFs per nucleotide substitution. An ORF is on average gained or lost at every 3.5 substitutions. We considered that iORFs with no detected ancestral homologs appeared on terminal branches. Among them, 91 to 93 % are present only in one lineage, which is consistent with the expected

conservation pattern for recently emerging iORFs (Table 1). The absence of ancestral homologs for the remaining 7 to 9% iORFs present in more than one lineage can be due to convergence on terminal branches, made possible by the relatively high turnover rate. Convergence events may particularly occur if two lineages acquire independently small indels, not necessarily at the same position but in the same iORF, leading to the same frameshift and resulting in stop codon changes. To estimate the expected frequency of such convergence events, we counted the number of iORFs for which the observed SNPs or indels lead to convergence in the alignment of 133 iORFs present in both *SpB* and *SpC* lineages, with no detected ancestors. We observed that ~30 % of patterns of convergence were indeed due of convergent mutations. Finally, regions with a higher rate of evolution may more likely lead to ancestral sequence reconstruction errors and to a small overestimation of the gain rate but this effect should be negligible because of the small number of iORFs with ambiguous age estimation.

Supplemental Methods

Ribosome profiling and mRNA sequencing libraries

Polysome extract preparation

Ribosome profiling and mRNA sequencing experiments were conducted with the strains YPS128 (*S. cerevisiae*) (Sniegowski et al. 2002), YPS744 (*S. paradoxus*), MSH604 (*S. paradoxus*) and MSH587 (*S. paradoxus*) belonging respectively to groups *SpA*, *SpB* and *SpC* according to Leducq et al. (2016). All strains were diploid. For *S. paradoxus*, we constructed homozygous diploids from haploid heterothallic strains containing a resistance cassette (Nourseothricine or Hygromycin B depending of the mating type) at the *HO* locus. Constructions and crosses were performed according to the protocol described in Leducq et al. (2016). Resulting diploid cells containing the

two resistance cassettes were selected on solid YPD (Yeast Peptone Dextrose) medium containing 100 ug/ml of Nourseothricine and 250 ug/ml of Hygromycin B.

Strains were grown overnight in 50 mL of SOE (Synthetic Oak Exudate) medium (Murphy et al. 2006), at 30°C with shaking at 250 rpm. These pre-cultures were used to inoculate a 30°C pre-warm 750 mL SOE medium at an initial OD₆₀₀ of ~0.03, and grown to an OD₆₀₀ between 0.6 to 0.7, at 30°C and shaking at 250 rpm. We choose the SOE medium to be closed to natural conditions in which *de novo* genes could emerge in wild yeast strains. Cultures were treated with cycloheximide (50 ug/mL final) for 5 minutes and cells were rapidly collected by vacuum filtration using a 90 mm cellulose nitrate filter with a 0.45 mm pore size and a fritted glass support. Cells were resuspended on ice in 2.5 mL of polysome lysis Buffer (10 mM Tris-HCL pH 7.4, 100 mM NaCl, 30 mM MgCl₂, 50 ug/mL cycloheximide). The slurry was then pipetted and frozen by fractions of ~ 20 ul in liquid nitrogen, and stored at -80°C. The resulting cryogenized mix was grinded in a MixerMill 400 (RETSCH) for 15 cycles of 2mn at 30 Hz, with chilling in liquid nitrogen between each cycle. The powder was gently thawed in the open grinding chamber at room temperature to collect the lysate which was cleared by two rounds of centrifugation for 5 minutes at 3000xg at 4°C, followed by one round of high speed centrifugation for 10 min at 20,000xg at 4°C. The middle layer was quantified by OD₂₆₀ measurement via Nanodrop and samples above 200 OD₂₆₀/mL were diluted to ~200 OD₂₆₀/mL in lysis buffer. Cell lysates were divided into 250 ul aliquots of 30 to 50 OD₂₆₀ each, that were flash frozen in liquid nitrogen and stored at -80°C. For each lysate, one aliquot of 250 ul was conserved for direct total mRNA extraction, the other aliquots were pooled for ribosome footprint isolation.

Isolation and purification of ribosome footprints

Cell lysate, corresponding to 50 to 100 OD₂₆₀, were digested with 15 U of RNase I (Ambion) per OD₂₆₀ for 60 minutes at 25 °C with shaking. The digestion was stopped by adding 200 U of Superase-in (AMBION). The digested products were loaded on a

24% sucrose cushion (50 mM Tris-acetate pH 7.6, 50 mM NH₄Cl, 12 mM MgCl₂, 1 mM DTT) and centrifuged at 4°C 100,000 rpm in a TLa110 rotor for 2h15. Pellets were washed two times with lysis buffer and resuspended in 500 uL of polysome lysis buffer. The extract was treated with DNase I using the manufacturer's instructions (Truseq Ribo profile illumina kit for yeast). RNA was then extracted using acid-phenol-chloroform extraction protocol, and precipitate overnight at -20 °C with 0.1 Volume of sodium acetate 3M, pH 5.2 and 3 volumes of EtOH 100%. Samples were centrifuged at 4°C for 20 minutes at 10,000 g and pellets were resuspended in 75 ul of RNAse free H₂O supplied with 1 U/uL of Superase-in (AMBION). RNA concentration was measured at 260 nm, resulting in a final amount of ~300 to 1000 ng of digested RNA. RNA fragments were separated by electrophoresis on a denaturing 17% PAGE gel with heating at 60 °C at 200V for ~8 hours. A mix of 28 and 34 nt RNA markers, oNTI199 and oNTI34ARN (Ingolia et al. 2012), was loaded at both extremities of the gel. The gel was stained with SYBR Gold according to the manufacturer's instructions and the region corresponding to the 28 marker was excised. Gel slices were disrupted through needle holes in a 0.5 mL centrifuge tubes nested in a 1.5 mL tube by maximum speed centrifugation. RNA was eluted overnight at 4°C with gentle rotation in an elution buffer (300 mM NaOAc pH 5.5, 1 mM EDTA). The slurry was loaded on SpinX cellulose acetate filter to recover the eluted RNA cleared of gel fragments. The RNA was precipitate overnight at -20°C in ethanol with 0.3 M sodium acetate and 20 ug of glycogen. Samples were centrifuged at 4°C for 30 minutes at maximum speed and pellets were resuspended in 25 ul of nuclease-free water supplemented with 0.1 U/mL of Superase-in (AMBION). These samples contain purified ribosome footprints. Total mRNA was extracted using the same acid-phenol-chloroform extraction protocol as for ribosome footprints samples. Purified ribosome footprints and total RNA were then quantified by fluorescence (Quant-it RNA assay kit, thermofisher) and stored at -20°C.

Library preparation

The rRNA was depleted in purified ribosome footprints and total mRNA samples using the Ribo-Zero Gold rRNA Removal Kit for yeast (Illumina) according to the manufacturer's instructions. Ribo-Zero treated RNAs were then purified by overnight ethanol precipitation. Ribosome profiling and total mRNA libraries were constructed using the TruSeq Ribo Profile kit for yeast (Illumina), using manufacturer's instructions starting from Fragmentation and end repair step. Circularized cDNA templates were amplified by 11 cycles of PCR using Phusion-polymerase (New England Biolabs), with primers incorporating barcoded Illumina TruSeq library sequences, according to TruSeq Ribo Profile kit for yeast (Illumina). The resulting PCR products were loaded onto a 8% native polyacrylamide gel in TBE and purified using the PCR purification protocol provided in the TruSeq Ribo Profile kit for yeast (Illumine). The quality and size of the purified PCR products were assessed using an Agilent HS bioanalyzer. Libraries were quantified by fluorescence using the Quant-iT PicoGreen dsDNA Assay Kit (ThermoFisher). The 8 total RNA libraries were pooled in one bulk for sequencing. RPF libraries were pooled in 2 bulks, one for the first replicate of the 4 strains, and a bulk for the second replicate of the 4 strains. Libraries were sequenced on an Illumina HiSeq 2500 at The Genome Quebec Innovation Center. The total RNA bulk was loaded onto 5 lanes and RPF bulks were loaded onto 4 lanes each.

Bibliography

Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charlotteaux B, Hidalgo CA, Barbette J, Santhanam B et al. 2012. Proto-genes and de novo gene birth. *Nature* **487**: 370-374.

Charron G, Leducq JB, Landry CR. 2014. Chromosomal variation segregates within incipient species and correlates with reproductive isolation. *Mol Ecol* **23**: 4362-4372.

Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. 2012. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc* **7**: 1534-1550.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**: 241-254.

Leducq JB, Nielly-Thibault L, Charron G, Eberlein C, Verta JP, Samani P, Sylvester K, Hittinger CT, Bell G, Landry CR. 2016. Speciation driven by hybridization and chromosomal plasticity in a wild yeast. *Nat Microbiol* **1**: 15003.

Murphy HA, Kuehne HA, Francis CA, Sniegowski PD. 2006. Mate choice assays and mating propensity differences in natural yeast populations. *Biol Lett* **2**: 553-556.

Peter J, De Chiara M, Friedrich A, Yue JX, Pflieger D, Bergstrom A, Sigwalt A, Barre B, Freel K, Llored A et al. 2018. Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* **556**: 339-344.

Sniegowski PD, Dombrowski PG, Fingerman E. 2002. *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* coexist in a natural woodland site in North America and display different levels of reproductive isolation from European conspecifics. *FEMS Yeast Res* **1**: 299-306.